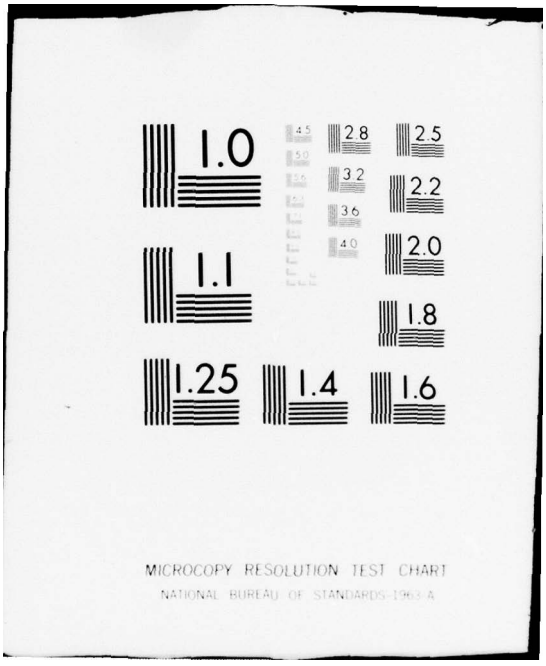
AD-A057 958 DAVID W TAYLOR NAVAL SHIP RESEARCH AND DEVELOPMENT CE--ETC F/G 12/1
MULTIPLE LINEAR REGRESSION.(U)
AUG 78 G R HUMFELD

UNCLASSIFIED DTNSRDC-78/060 NL

END
DATE FILMED
10-78
DDC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

LEVEL II

DTNSRDC-78/060

DAVID W. TAYLOR NAVAL SHIP RESEARCH AND DEVELOPMENT CENTER

Bethesda, Md. 20084

MULTIPLE LINEAR REGRESSION

by

George R. Humfeld

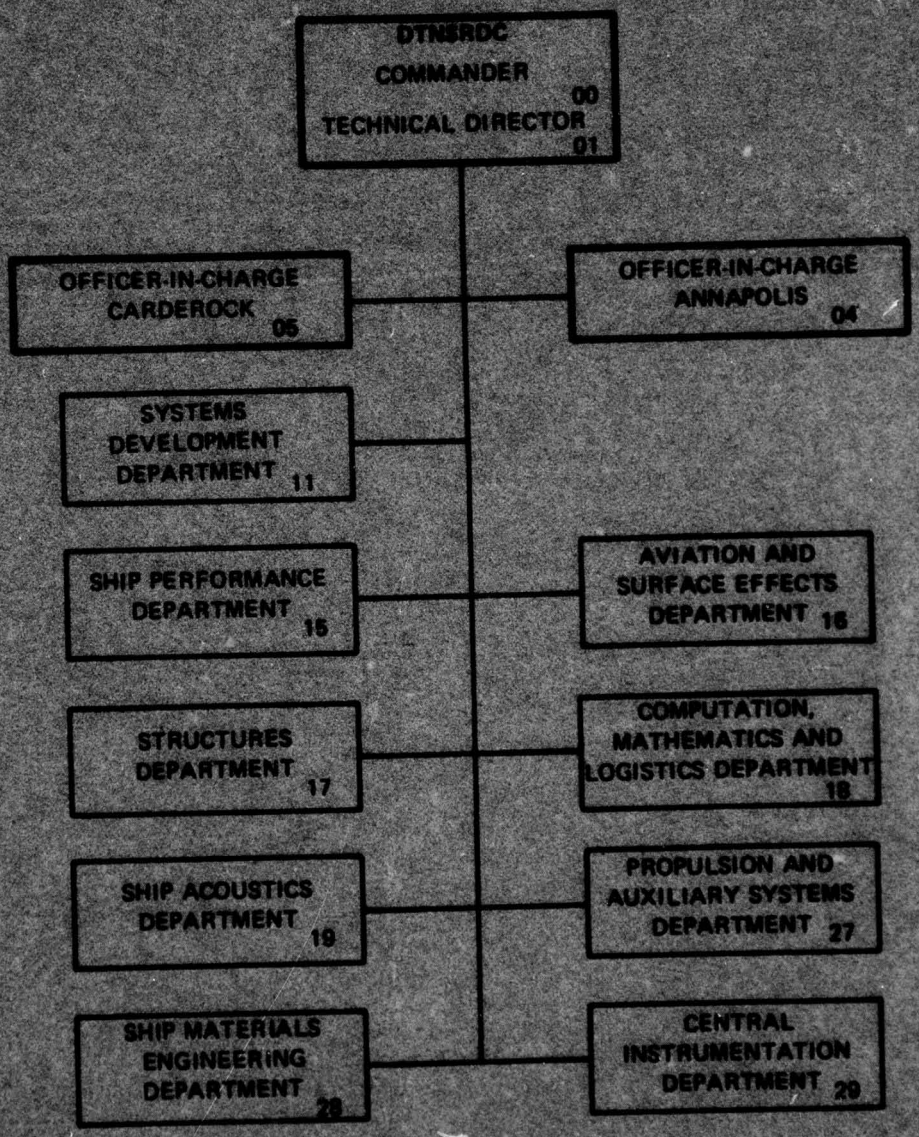
COMPUTATION, MATHEMATICS, AND LOGISTICS DEPARTMENT
RESEARCH AND DEVELOPMENT REPORT

August 1978

DTNSRDC-78/060

MULTIPLE LINEAR REGRESSION

MAJOR DTNSRDC ORGANIZATIONAL COMPONENTS



UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: DTNSRDC-78/060	2. GOVT ACCESSION NO. / 3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): MULTIPLE LINEAR REGRESSION	5. TYPE OF REPORT & PERIOD COVERED: Final Rept.
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): George R. Humfeld	8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS: David W. Taylor Naval Ship Research and Development Center, Bethesda, Maryland 20084	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: (See reverse side)
11. CONTROLLING OFFICE NAME AND ADDRESS	12. REPORT DATE: August 1978
	13. NUMBER OF PAGES: 34
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): 34 p.	15. SECURITY CLASS. (of this report): UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Multiple Linear Regression
Data Analysis
Estimation
Prediction
Operations Research

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Multiple linear regression theory provides an estimated covariance structure for the estimates of the parameters of the linear function based on given data. However, when the deviation form is used to calculate these parameter estimates, the portions of this matrix which involve the constant term are generally missing. This report presents equations which

(Continued on reverse side)

next page

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

387 682

(Block 10)

Program Element 60000N
Task Area OMN
Work Unit 1-1870-003

(Block 20 continued)

can be used to calculate these missing covariances from quantities which are generally available.

Most standard regression references discuss calculation of confidence limits for point estimates of the dependent variable when these point estimates are calculated from the regression equation. This report presents equations for similar limits for the independent variables, again from quantities generally available when a deviation-form routine is used. A different interpretation is suggested for these limits than is seen in the references.

A numerical example is provided.

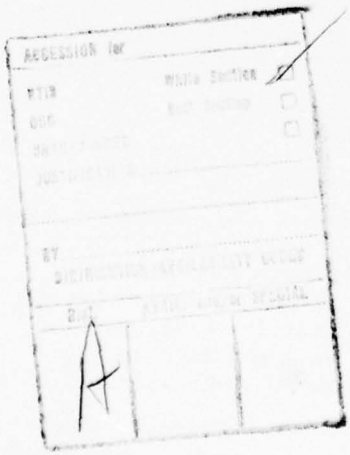


TABLE OF CONTENTS

LIST OF TABLES

ABSTRACT

Multiple linear regression theory provides an estimated covariance structure for the estimates of the parameters of the linear function based on given data. However, when the deviation form is used to calculate these parameter estimates, the portions of this matrix which involve the constant term are generally missing. This report presents equations which can be used to calculate these missing covariances from quantities which are generally available.

Most standard regression references discuss calculation of confidence limits for point estimates of the dependent variable when these point estimates are calculated from the regression equation. This report presents equations for similar limits for the independent variables, again from quantities generally available when a deviation-form routine is used. A different interpretation is suggested for these limits than is seen in the references.

A numerical example is provided.

ADMINISTRATIVE INFORMATION

This report is a result of work performed under Program Element 60000N, Task Area OMN, and Work Unit 1-1870-003.

1. INTRODUCTION

Multiple linear regression is a method of determining a linear relationship between a dependent variable and a collection of independent variables. The dependent variable is assumed to be equal to the sum of a linear function of the independent variables and a random variable which has zero mean and unknown variance. Multiple linear regression provides the (least squares) best estimates of the parameters of the linear function based on given data. In addition, it also can provide indications of how well the calculated linear function fits the data, how much the parameter estimates might vary from the "true" values, and how much point estimates obtained by using the calculated regression equation might vary from the "true" values.

The calculations involved in determining the parameter estimates and other quantities may be provided in either of two forms. The standard form determines a constant term and a coefficient for each of the independent

variables. The deviation form replaces each data value for each variable with its deviation from the sample mean (for that variable) and determines only the coefficients. The constant term is readily calculated from the means of the variables and the other parameter estimates. Other quantities, although also available by calculation, are generally ignored in a discussion of the deviation form. It is the purpose of this report to develop expressions for these quantities.

Many books and papers have been written on multiple linear regression. The reader is assumed to have been familiar at one time with the techniques involved. However, sufficient review is given that this familiarity need not be recent. Some of the better known facts will be stated without proof in this report. For background the reader is referred to Acton,[1*] Draper and Smith[2] or Johnston.[3] Johnston is the primary reference used by the author.

Multiple linear regression theory provides an estimated covariance structure for the parameter estimates. However, when the deviation form is used to calculate these parameter estimates, the portions of this matrix which involve the constant term are generally missing. This is the case, for example, when the International Mathematical and Statistical Libraries (IMSL) routines[4] RLSTEP and RLFORC are used in a stepwise multiple linear regression application. This report derives equations which can be used to calculate these missing covariances from quantities which are generally available: the sum of the squared residuals, the means of the variables, and the remainder of the estimated covariance matrix.

Most standard regression references discuss calculation of confidence limits for point estimates of the dependent variable when these point estimates are calculated from the regression equation. Some (for example, Acton[1]) also discuss similar limits for similarly obtained point estimates for the independent variables. However, these discussions generally are restricted to cases in which there is a single independent variable. Also, the deviation forms are not considered in this context. Equations are derived in this report for calculation of such limits in cases in which

---

*A complete listing of references is given on page 29.

there is more than one independent variable and the regression equation has been calculated using the deviation form. In this report a different interpretation is suggested for these limits than is seen in the references.

Section 2 reviews the assumptions involved in use of a multiple linear regression procedure. The notation used throughout the report is also introduced in this section.

Section 3 provides a basic review of multiple linear regression and states the pertinent equations used in calculating the parameter estimates and other quantities when the standard form is used.

Section 4 develops similar equations using the deviation form. Equations are included for the complete estimated covariance matrix of the parameter estimates. A familarity with some simple matrix operations is necessary for a proper understanding of this section.

Section 5 specifies, step by step, the numerical procedures to be followed in the application of multiple linear regression as described in Section 4. In particular, the matrix equations of Section 4 are rewritten in a nonmatrix form which can be used in a computer program. The limits on predicted values of an independent variable are discussed in this section.

A numerical example is provided in Section 6. This example follows the step-by-step procedure given in Section 5. It is of sufficient complexity to exemplify each step and yet of sufficient simplicity to allow hand calculation.

A final section provides a summary.

2. ASSUMPTIONS AND NOTATION

Given a dependent variable Y and k - 1 independent variables X_2, X_3, ..., X_k, multiple linear regression assumes a model of the form

$$Y = b_1 + b_2X_2 + b_3X_3 + \ldots + b_kX_k + u \qquad (1)$$

where the parameters b_1, b_2, ..., b_k are fixed constants and where u is a random variable with zero mean and constant, but unknown, variance w^2. Generally, Y must be a random variable since u is, but the independent variables are assumed to be nonrandom. Note that it is not necessary at this point to assume that the random errors u are normally distributed.

The assumed nonrandomness of the independent variables means that the values of these variables have been measured accurately and are therefore known exactly. Strictly speaking, randomness in the independent variables (for example, inaccuracies in measuring and recording their values) violates this assumption. However, this fact is usually ignored when regression is applied, the prevailing feeling being that the (hopefully small) randomness in the independent variables can be considered as a part of the randomness in u. The robustness of the method often provides useful results in such cases.

In general, column vectors are favored over row vectors in this report. The transpose of a vector or matrix is indicated by a prime. For convenience, each vector will be introduced in terms of its transpose. Thus, $\underline{b}$, the vector of parameters, is introduced by $\underline{b}' = (b_1, b_2, \ldots, b_k)$.

The vector, all of whose components are zero, is denoted by $\underline{0}$. The vector, all of whose components are one, is denoted by $\underline{1}$. When used, the size of these vectors is clear from the context. Similarly, the dimension of any identity matrix, I, is clear from its use.

The data are assumed to be arranged in k-tuples, called data points. The $i^{\underline{th}}$ data point is $(X_{2i}, X_{3i}, \ldots, X_{ki}, Y_i)$, where Y_i is the $i^{\underline{th}}$ observed value of the dependent variable and X_{ji} is the $i^{\underline{th}}$ observed value of the $j^{\underline{th}}$ independent variable. There are assumed to be n data points. The values in each point are assumed to be associated according to Equation (1), so that

$$Y_i = b_1 + b_2 X_{2i} + \ldots + b_k X_{ki} + u_i \qquad (2)$$

where u_i is some unknown value of the random variable u. In matrix form Equation (2) may be written

$$\underline{Y} = X\underline{b} + \underline{u} \tag{3}$$

where $\underline{Y}' = (Y_1, Y_2, \ldots, Y_n)$, $\underline{b}' = (b_1, b_2, \ldots, b_k)$, $\underline{u}' = (u_1, u_2, \ldots, u_n)$, and

$$X = \begin{bmatrix} 1 & X_{21} & X_{31} & \cdots & X_{k1} \\ 1 & X_{22} & X_{32} & \cdots & X_{k2} \\ . & . & . & \cdots & . \\ . & . & . & \cdots & . \\ . & . & . & \cdots & . \\ 1 & X_{2n} & X_{3n} & \cdots & X_{kn} \end{bmatrix} \tag{4}$$

It is further assumed that no values are missing. That is, a data point is not used in the regression unless all variables in it have values.

The sample mean of a variable is denoted by a bar above the variable name. Thus,

$$\bar{Y} = \left(\sum_{i=1}^{n} Y_i \right) / n = \underline{1}'\underline{Y}/n$$

$$\bar{X}_j = \left(\sum_{i=1}^{n} X_{ji} \right) / n$$

The deviation of a variable from its sample mean is denoted by replacing the upper case symbol by the corresponding lower case symbol:

$$y_i = Y_i - \bar{Y}$$

$$x_{ji} = X_{ji} - \bar{X}_j$$

The sample variances, covariances, and correlations are easily expressed in terms of these deviations. For example:

$$\text{var}(X_j) = \left(\sum_{i=1}^{n} x_{ji}^2 \right) / (n-1)$$

$$\text{cov}(X_j, X_h) = \left(\sum_{i=1}^{n} x_{ji} x_{hi} \right) / (n-1)$$

$$\text{corr}(X_j, X_h) = \left(\sum_{i=1}^{n} x_{ji} x_{hi} \right) \Bigg/ \left(\sum_{i=1}^{n} x_{ji}^2 \right)^{1/2} \left(\sum_{i=1}^{n} x_{hi}^2 \right)^{1/2}$$

For each j = 1, 2, ..., k, multiple linear regression finds an estimate B_j for the value of the parameter b_j in Equation (1). The result, then, is the regression equation:

$$\tilde{Y} = B_1 + B_2 X_2 + B_3 X_3 + \ldots + B_k X_k \qquad (5)$$

which can be used to arrive at an estimated value for any variable given values for the others. (Caution: No guarantee has yet been given concerning the accuracy of such estimates.) In particular, for i = 1, 2, ..., n,

$$\tilde{Y}_i = B_1 + B_2 X_{2i} + B_3 X_{3i} + \ldots + B_k X_{ki} \qquad (6)$$

is the estimated, or predicted, value of Y in the $i^{\underline{th}}$ data point. This value can be compared to the actual value, Y_i. The difference

$$e_i = Y_i - \tilde{Y}_i \tag{7}$$

is called the residual and represents what must be added to the predicted value to get the actual value. In matrix form the vector of residuals, $\underline{e} = (e_1, e_2, \ldots e_n)$, is given by

$$\underline{e} = \underline{Y} - \tilde{\underline{Y}} = \underline{Y} - X\underline{B} \tag{8}$$

Note from Equation (7) that e_i is an estimate of the value of u_i. This fact is used in the next section to arrive at an estimate for w^2.

The matrix Equations (3) and (8) provide a starting point for an elegant development of the multiple linear regression equations given without proof in the next section. The interested reader is referred to Johnston,[3] Chapter 5.

In addition to the data matrix in standard form given in Equation (4), the data matrix in deviation form:

$$x = \begin{bmatrix} x_{21} & x_{31} & \cdots & x_{k1} \\ x_{22} & x_{32} & \cdots & x_{k2} \\ . & . & \cdots & . \\ . & . & \cdots & . \\ . & . & \cdots & . \\ x_{2n} & x_{3n} & \cdots & x_{kn} \end{bmatrix} \tag{9}$$

will also be used.

In addition to the assumptions given earlier in this section, it is necessary to assume that the random variable u is normally distributed in order to derive confidence intervals. For other results, however, this

assumption is not needed. On the other hand, it is always necessary to assume independence of u_i and u_h for i not equal to h. Relaxation of many of the assumptions stated here is considered in Johnston.[3]

3. STANDARD REGRESSION EQUATIONS

In this section the standard multiple linear regression equations are presented in matrix form. Since development of these equations is contained in any good book on regression analysis, it is not repeated here. Recall from the previous section that X is the data matrix in standard form, $\underline{Y}$ is the vector of sample values of the dependent variable, $\underline{b}$ is the vector of regression parameters, $\underline{B}$ is the vector of parameter estimates, w^2 is the variance of the random variable u (see Equation (1)), and $\underline{e}$ is the vector of residuals. Also used in this section is the vector of deviations of the value of the dependent variable about its mean: $\underline{y} = (y_1, y_2, \ldots, y_n)$.

The objective of multiple linear regression is to determine the vector $\underline{B}$ which will minimize $\underline{e}'\underline{e}$, the sum of the squared residuals. The vector $\underline{B}$ which accomplishes this objective is given by

$$\underline{B} = (X'X)^{-1}X'\underline{Y} \tag{10}$$

The matrix X'X is a square symmetric matrix called the information matrix. Note that each B_j is a linear combination of the random variables $Y_1, Y_2, \ldots, Y_n$. (The elements of X were assumed to be nonrandom.) From the assumptions imposed upon u, it is determined that

$$E(\underline{B}) = \underline{b} \tag{11}$$

$$\text{var}(\underline{B}) = w^2(X'X)^{-1} \tag{12}$$

So $\underline{B}$ is an unbiased estimator of $\underline{b}$ (that is, B_j is an unbiased estimator of b_j for each j = 1, 2, ..., k). The Gauss-Markov Theorem on least

squares indicates that $\underline{B}$ is the best linear unbiased estimator of $\underline{b}$. Although Equation (12) gives the covariance structure of $\underline{B}$, w^2 is unknown and var($\underline{B}$) cannot actually be determined. However, with the choice of $\underline{B}$ given in Equation (10), the expected value of the sum of the squared residuals is given by

$$E(\underline{e}'\underline{e}) = (n-k)w^2 \tag{13}$$

Thus, w^2 can be approximated

$$w^2 \doteq v^2 = (\underline{e}'\underline{e})/(n-k) \tag{14}$$

From the approximation of Equation (12), an estimated covariance matrix for $\underline{B}$ is found to be

$$\text{var}(\underline{B}) \doteq v^2(X'X)^{-1} = (\underline{e}'\underline{e})(X'X)^{-1}/(n-k) \tag{15}$$

The coefficient of multiple correlation, R^2, often used as a measure of the goodness of fit of the regression equation to the given data, is calculated from

$$R^2 = 1 - (\underline{e}'\underline{e})/(\underline{y}'\underline{y}) \tag{16}$$

The regression equation, Equation (5), can be used to determine a predicted value for the dependent variable given values for the independent variables. On the assumption that u (and therefore Y and $\underline{B}$) is normally distributed, confidence limits may be placed around this prediction. For example, suppose that it is desired to determine such confidence limits for Y when $X_j = Z_j$ for j = 2, 3, ..., k. For $\underline{Z}' = (1, Z_2, Z_3, \ldots, Z_k)$, $\widetilde{Y}$ is normally distributed and

$$E(\widetilde{Y}) = \underline{Z}'\underline{b} \tag{17}$$

$$\text{var}(E(\widetilde{Y})) = w^2 \underline{Z}'(X'X)^{-1}\underline{Z} \tag{18}$$

$$\text{var}(\widetilde{Y}) = w^2(1 + \underline{Z}'(X'X)^{-1}\underline{Z}) \tag{19}$$

However, $\underline{e}'\underline{e}$ has a chi-squared distribution with n-k degrees of freedom and is independent of $\underline{Z}'\underline{B}$. The confidence limits are found, by shifting in the usual way to a Student's t distribution, to be, for $E(\widetilde{Y})$,

$$\underline{Z}'\underline{B} \pm tv\sqrt{\underline{Z}'(X'X)^{-1}\underline{Z}} \tag{20}$$

and, for $\widetilde{Y}$,

$$\underline{Z}'\underline{B} \pm tv\sqrt{1 + \underline{Z}'(X'X)^{-1}\underline{Z}} \tag{21}$$

where t is a critical point from a t distribution with n-k degrees of freedom. If r is the desired significance level, then t is the (1 - r/2) critical point. (For example, for 90 percent confidence limits the significance level is r = 0.1, and t would be found in the 0.95 column of the table of t distribution points.)

The regression equation can also be used to calculate a predicted value for one of the independent variables given values for the dependent variable and each of the other independent variables. Limits similar to confidence limits, can also be calculated in this case. These limits will be discussed in Section 5.

4. REGRESSION EQUATIONS IN DEVIATION FORM

In this section the deviation form equivalents for Equations (10), (12), (15), (17), (18), (19), (20), and (21) are presented. As noted in the introduction, the deviation form equivalent of Equation (15) does not contain the portion of the estimated covariance matrix of the parameter estimates which deals with the constant term. Equations for these covariances are derived in this section.

First, consider the structure of X'X and x'x. From Equation (4), the definition of X, note that

$$X'X = \begin{bmatrix} n & n\bar{\underline{X}}' \\ n\bar{\underline{X}} & S \end{bmatrix} \tag{22}$$

where $\bar{\underline{X}} = (\bar{X}_2, \bar{X}_3, \ldots, \bar{X}_k)$ and S is the $(k-1)\times(k-1)$ matrix having at the intersection of its $(s-1)^{\underline{th}}$ row and its $(t-1)^{\underline{th}}$ column the element

$$S_{s-1,t-1} = \sum_{i=1}^{n} X_{si}X_{ti} \tag{23}$$

From Equation (9), the definition of x, it is seen that the element at the intersection of the $(s-1)^{\underline{th}}$ row and the $(t-1)^{\underline{th}}$ column is given by

$$(x'x)_{s-1,t-1} = \sum_{i=1}^{n} x_{si}x_{ti}$$

$$= \sum_{i=1}^{n} X_{si}X_{ti} - n\bar{X}_s\bar{X}_t \tag{24}$$

From a comparison of Equations (23) and (24), the matrix form for Equation (24) is found to be

$$x'x = S - n\bar{\underline{X}}\bar{\underline{X}}' \tag{25}$$

The relationship between the inverses of X'X and x'x is clarified by an examination of the matrix

$$Q = \begin{bmatrix} 1/n & \underline{0}' \\ -\underline{\bar{X}} & I \end{bmatrix} \tag{26}$$

Then,

$$Q(X'X) = \begin{bmatrix} 1 & \underline{\bar{X}}' \\ \underline{0} & x'x \end{bmatrix} \tag{27}$$

Inversion of Equation (27) shows that

$$(X'X)^{-1}Q^{-1} = \begin{bmatrix} 1 & -\underline{\bar{X}}'(x'x)^{-1} \\ \underline{0} & (x'x)^{-1} \end{bmatrix} \tag{28}$$

When both sides of this last equation are multiplied on the right by Q as defined in Equation (26),

$$(X'X)^{-1} = \begin{bmatrix} 1/n + \underline{\bar{X}}'(x'x)^{-1}\underline{\bar{X}} & -\underline{\bar{X}}'(x'x)^{-1} \\ -(x'x)^{-1}\underline{\bar{X}} & (x'x)^{-1} \end{bmatrix} \tag{29}$$

Consideration is next given to the form of $X'\underline{Y}$ and $x'\underline{y}$. If X. is the submatrix of X found by eliminating the first column of X (consisting of all ones), it is found that

$$x'\underline{y} = X.'\underline{Y} - n\overline{Y}\underline{\bar{X}} \tag{30}$$

This is the matrix equivalent of

$$\sum_{i=1}^{n} x_{ji}y_i = \sum_{i=1}^{n} X_{ji}Y_i - n\bar{X}_j\bar{Y} \tag{31}$$

It is also found that

$$X'\underline{Y} = \begin{pmatrix} n\bar{Y} \\ X_{\cdot}'\underline{Y} \end{pmatrix} \tag{32}$$

The deviation form equivalent of Equation (10) is

$$\underline{B}_{\cdot} = (x'x)^{-1}x'\underline{y} \tag{33}$$

From Equations (10), (29), (30), (32), and (33), it is determined that

$$\underline{B} = \begin{pmatrix} \bar{Y} + n\bar{Y}\underline{\bar{X}}'(x'x)^{-1}\underline{\bar{X}} - \underline{\bar{X}}'(x'x)^{-1}X_{\cdot}'\underline{Y} \\ -n\bar{Y}(x'x)^{-1}\underline{\bar{X}} + (x'x)^{-1}X_{\cdot}'\underline{Y} \end{pmatrix}$$

$$= \begin{pmatrix} \bar{Y} - \underline{\bar{X}}'\underline{B}_{\cdot} \\ \underline{B}_{\cdot} \end{pmatrix} \tag{34}$$

That is, $\underline{B}_{\cdot}$ is the subvector of $\underline{B}$ found by eliminating B_1. Furthermore, B_1 may also be found from Equation (34):

$$B_1 = \bar{Y} - \underline{\bar{X}}'\underline{B}_{\cdot} \tag{35}$$

Note that Equations (15) and (29) yield the estimated covariance matrix for $\underline{B}$ in terms of x'x and the sample means:

$$\text{var}(\underline{B}_{\cdot}) = w^2(x'x)^{-1}$$

$$\doteq (e'e)(x'x)^{-1}/(n-k) \tag{36}$$

$$\text{cov}(\underline{B}.,B_1) = -w^2\bar{\underline{X}}'(x'x)^{-1}$$

$$\doteq -(e'e)\bar{\underline{X}}'(x'x)^{-1}/(n-k) \qquad (37)$$

$$\text{var}(B_1) = w^2(1/n + \bar{\underline{X}}'(x'x)^{-1}\bar{\underline{X}})$$

$$\doteq (e'e)(1/n + \bar{\underline{X}}'(x'x)^{-1}\bar{\underline{X}})/(n-k) \qquad (38)$$

Finally, consideration is given to estimation in the deviation context. Since all the elements of $\underline{B}$ may be calculated from Equations (33) and (35), point estimates may be calculated from the regression equation, Equation (5), or from the deviation form of Equation (5), which is:

$$\tilde{y} = B_2x_2 + B_3x_3 + \ldots + B_kx_k \qquad (39)$$

As in the preceding section, suppose that confidence limits for Y are desired when $X_j = Z_j$, for j = 2, 3, ..., k. If $\underline{Z}.' = (Z_2,Z_3,\ldots,Z_k)$ and $\underline{z}' = \underline{Z}! - \underline{X}' = (z_2,\ldots,z_k)$, where $z_j = Z_j - \bar{X}_j$, for each j, then

$$\underline{z}'(x'x)^{-1}\underline{z} = \underline{Z}.'(x'x)^{-1}\underline{Z}. - \underline{Z}.'(x'x)^{-1}\bar{\underline{X}}$$

$$- \bar{\underline{X}}'(x'x)^{-1}\underline{Z}. + \bar{\underline{X}}'(x'x)\bar{\underline{X}} \qquad (40)$$

Thus, Equation (29) implies that

$$\underline{Z}'(X'X)^{-1}\underline{Z} = 1/n + \underline{z}'(x'x)^{-1}\underline{z} \qquad (41)$$

When this result is used in Equations (18) through (21), the deviation equivalents are:

$$\text{var}(E(\tilde{y})) = w^2(1/n + \underline{z}'(x'x)^{-1}\underline{z}) \qquad (42)$$

$$\text{var}(\tilde{y}) = w^2[(n+1)/n + \underline{z}'(x'x)^{-1}\underline{z}] \tag{43}$$

$$\underline{z}'\underline{B}. \pm tv\sqrt{1/n + \underline{z}'(x'x)^{-1}\underline{z}} \tag{44}$$

$$\underline{z}'\underline{B}. \pm tv\sqrt{(n+1)/n + \underline{z}'(x'x)^{-1}\underline{z}} \tag{45}$$

where Equations (44) and (45) are confidence limits for $E(\tilde{y})$ and $\tilde{y}$, respectively. As in Section 3, v is as given in Equation (14) and t is a critical point from a t distribution with n-k degrees of freedom. The confidence limits for $\tilde{Y}$ may be obtained from Equation (45) by adding the mean $\bar{Y}$ to the limits for $\tilde{y}$. Equivalent to Equation (17) is

$$E(\tilde{y}) = \underline{z}'b. \tag{46}$$

where $\underline{b}.' = (b_2, b_3, \ldots, b_k)$.

In the next section some of the matrix equations of this section are put into a form more amenable to numerical calculation.

5. NUMERICAL PROCEDURES

Suppose a set of n data points is available for performance of a multiple linear regression. This section sets forth the steps of the numerical procedure to be used in accomplishing this regression and determining the other information discussed in earlier sections. The deviation form equations of the preceding section are used here.

The first step is to develop the matrix x'x and the vector x'$\underline{y}$. This may be accomplished with a single pass through the data by accumulating

$$\text{(a)} \quad \sum_{i=1}^{n} X_{ji}$$

$$\text{(b)} \quad \sum_{i=1}^{n} Y_i$$

$$(c) \quad \sum_{i=1}^{n} X_{ji} X_{mi}$$

$$(d) \quad \sum_{i=1}^{n} X_{ji} Y_i$$

for each j,m = 2, 3, ..., k. The means $\bar{X}_j$ and $\bar{Y}$ are found by dividing (a) and (b), respectively, by n. These means are then used with (c) to determine the elements of x'x from Equation (24), and with (d) to determine the elements of $x'\underline{y}$ from

$$(x'\underline{y})_{s-1} = \sum_{i=1}^{n} x_{si} y_i$$

$$= \sum_{i=1}^{n} X_{si} Y_i - n\bar{X}_s \bar{Y} \qquad (47)$$

So that R^2 may be calculated from Equation (16), $\sum_{i=1}^{n} Y_i^2$ should also be accumulated and used to determine $\underline{y}'\underline{y}$ from

$$\underline{y}'\underline{y} = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2 \qquad (48)$$

Since the means will be required later, they should be retained. Dorn and McCracken[5] indicate in Sections 3.5 and 7.8 that this single pass method

may not be as accurate numerically as a two-pass method in which (a) and (b) are accumulated on the first pass and the sums (over i) of $x_{ji}x_{mi}$, $x_{ji}y_i$, and y_i^2 on the second pass.

The second step is to invert x'x. Most modern computers have reasonably accurate matrix inversion routines available. For convenience the element of $(x'x)^{-1}$ at the intersection of row s and column t will be denoted by $c_{s+1,t+1}$. That is,

$$(x'x)^{-1} = \begin{bmatrix} c_{22} & c_{23} & \cdots & c_{2k} \\ c_{32} & c_{33} & \cdots & c_{3k} \\ . & . & \cdots & . \\ . & . & \cdots & . \\ . & . & \cdots & . \\ c_{k2} & c_{k3} & \cdots & c_{kk} \end{bmatrix} \tag{49}$$

The third step is to determine the estimates of the regression parameters. For j = 2, 3, ..., k, these estimates are calculated (see Equation (33)) from

$$B_j = \sum_{m=2}^{k} c_{jm} \left(\sum_{i=1}^{n} x_{mi} y_i \right) \tag{50}$$

The value of B_1 may then be calculated (see Equation (35)) from

$$B_1 = \bar{Y} - \sum_{j=2}^{k} B_j \bar{X}_j \tag{51}$$

The regression equation, Equation (5), is now known and can be used to arrive at a point estimate for Y, given values $X_2 = Z_2$, $X_3 = Z_3$, ..., $X_k = Z_k$ for the independent variables. It can also be used to arrive at a point estimate for independent variable X_m given $Y = Y'$ and, for $j = 2, 3, \ldots, m-1, m+1, \ldots, k$, $X_j = Z_j$ from

$$\tilde{X}_m = \left(Y' - B_1 - \sum_{j=2}^{k}{}^{m} B_j Z_j \right) / B_m \tag{52}$$

where $\sum^{m}$ refers to the sum exclusive of the $m^{\underline{th}}$ term.

The next step is to calculate R^2 from Equation (16). Equations (5) and (7) may be used and a pass may be made through the data to calculate the residual associated with each data point. The sum of the squared residuals can then be accumulated and used in Equation (16) to calculate R^2. An alternative procedure, which does not necessitate calculation of the individual residuals, is based on the fact (see Johnston,[3] Equation (5.22)) that

$$\underline{e}'\underline{e} = \underline{y}'\underline{y} - \underline{B}.'x'\underline{y}$$

$$= \underline{y}'\underline{y} - \sum_{j=2}^{k} B_j \left(\sum_{i=1}^{n} x_{ji} y_i \right) \tag{53}$$

This leads to

$$R^2 = \left[\sum_{j=2}^{k} B_j \left(\sum_{i=1}^{n} x_{ji} y_i \right) \right] / \underline{y}'\underline{y} \tag{54}$$

In addition, Equation (53) can be used in Equation (14) to provide the following estimate of w^2, the variance of the dependent variable:

$$w^2 \doteq v^2$$

$$\doteq \left[\underline{y}'\underline{y} - \sum_{j=2}^{k} B_j \left(\sum_{i=1}^{n} x_{ji} y_i \right) \right] \Big/ (n-k) \qquad (55)$$

The fifth step is to calculate the estimated covariance structure of the parameter estimates from Equations (36), (37), and (38). From Equation (38) the estimated variance of B_1 is seen to be

$$\text{var}(B_1) \doteq v^2 \left(1/n + \sum_{s=2}^{k} \sum_{t=2}^{k} c_{st} \bar{X}_s \bar{X}_t \right) \qquad (56)$$

For j = 2, 3, ..., k, Equation (37) indicates that

$$\text{cov}(B_j, B_1) \doteq - v^2 \sum_{s=2}^{k} c_{js} \bar{X}_s \qquad (57)$$

Finally, for j,m = 2, 3, ..., k, Equation (36) yields

$$\text{cov}(B_j, B_m) \doteq v^2 c_{jm} \qquad (58)$$

For any given set of values, $X_2 = Z_2$, $X_3 = Z_3$, ..., $X_k = Z_k$, confidence limits at various levels of significance can be calculated for the dependent variable using Equation (45). For 100(1-r) percent confidence limits use

$$B_1 + \sum_{j=2}^{k} B_j Z_j$$

$$\pm\, tv\sqrt{(n+1)/n + \sum_{s=2}^{k}\sum_{t=2}^{k} c_{st}(Z_s - \bar{X}_s)(Z_t - \bar{X}_t)} \qquad (59)$$

where t is the (1 - r/2) critical point from a t distribution with n-k degrees of freedom. Note that the point estimate of the dependent variable given by Equation (5) is midway between these limits and that the length of the interval bounded by these limits is 2t times the square root of the estimated variance of the point estimate (see Equation (43)).

Converting Equation (59) to a probability statement produces a method of calculating limits, similar to confidence limits, for the predicted value of an independent variable, given values for the dependent variable and each of the other independent variables. Since the independent variables are not random (see Section 2), these limits are not truly confidence limits. However, they do have a potentially useful interpretation as a specification of the range of values for the independent variable for which a particular value lies within a confidence interval for the dependent variable.

Suppose that values have been specified for $X_2, X_3, \ldots, X_{m-1}, X_{m+1}, \ldots, X_k$ as above. Suppose further that it is desired to find the values of X_m for which $Y = Y'$ will lie in a 100(1-r) percent confidence interval for (the random variable) Y. For each value Z_m of X_m,

$$\text{prob}\left\{\left|Y - B_1 - \sum_{j=2}^{k} B_j Z_j\right| \leq tv\sqrt{(n+1)/n + \sum_{s=2}^{k}\sum_{t=2}^{k} c_{st}(Z_s - \bar{X}_s)(Z_t - \bar{X}_t)}\right\}$$

$$= 1 - r \qquad (60)$$

If the inequality within the braces in Equation (60) is solved for the unknown Z_m with $Y = Y'$, the values of X_m for which $Y = Y'$ lies in a 100(1-r) percent confidence interval for Y may be determined.

The result of such an algebraic procedure is

$$fZ_m^2 + gZ_m + h \leq 0 \tag{61}$$

where

$$f = B_m^2 - t^2 v^2 c_{mm} \tag{62}$$

$$g = -2t^2v^2 \left[\sum_{j=2}^{k\ m} c_{jm}(Z_j - \bar{X}_j) - c_{mm}\bar{X}_m \right] - 2B_m \left[Y' - B_1 - \sum_{j=2}^{k\ m} B_j Z_j \right] \tag{63}$$

$$h = \left[Y' - B_1 - \sum_{j=2}^{k\ m} B_j Z_j \right]^2 - t^2 v^2 \left[(n+1)/n + \sum_{s=2}^{k\ m} \sum_{t=2}^{k\ m} c_{st}(Z_s - \bar{X}_s)(Z_t - \bar{X}_t) - 2\bar{X}_m \sum_{j=2}^{k\ m} c_{mj}(Z_j - \bar{X}_j) + c_{mm}\bar{X}_m^2 \right] \tag{64}$$

The zeros of the function on the left side of Equation (61) can be determined using the quadratic formula. If this function has no real zeros, $Y = Y'$ lies in the confidence interval for all values of X_m (provided $X_1 = Z_1, \ldots, X_{m-1} = Z_{m-1}, X_{m+1} = Z_{m+1}, \ldots, X_k = Z_k$). Otherwise, the zeros may be denoted by Z_u and Z_e, Z_u being the larger of the two values. Then Equation (61) may be used to determine the values of X_m for which $Y = Y'$ lies in the confidence interval. These values are all those between Z_e and Z_u, if $\tilde{X}_m$ from Equation (52) is between Z_e and Z_u, and all values less than or equal to Z_e or larger than or equal to Z_u otherwise.

It should be remembered that the last few paragraphs depend on the assumption that the random variable u (and therefore the dependent variable) has a normal distribution, as well as the other assumptions stated in Section 2. Such normality assumptions imply normality of the residuals. If the residual is calculated for each data point, a standard normality test will indicate how valid such an assumption may be.

6. A NUMERICAL EXAMPLE

The example discussed in this section is discussed in some detail by Draper and Smith,[2] although they do not cover some of the details considered here. The development presented here follows the steps of the preceding section.

The data consists of $n = 13$ data points, each point consisting of a value of a dependent variable, Y, and each of two independent variables, X_2 and X_3. The data are listed in Table 1. Note, for example, that $X_{2,4} = 11$, $X_{3,6} = 55$, and $Y_7 = 102.7$.

TABLE 1 - DATA FOR NUMERICAL EXAMPLE

Data Point Number	X_2	X_3	Y
1	7	26	78.5
2	1	29	74.3
3	11	56	104.3
4	11	31	87.6
5	7	52	95.9
6	11	55	109.2
7	3	71	102.7
8	1	31	72.5
9	2	54	93.1
10	21	47	115.9
11	1	40	83.8
12	11	66	113.3
13	10	68	109.4

Step 1:

$$\sum_{i=1}^{13} X_{2i} = 97 \quad \bar{X}_2 = 7.461538 \quad \sum_{i=1}^{13} X_{2i}^2 = 1139$$

$$\sum_{i=1}^{13} X_{3i} = 626 \quad \bar{X}_3 = 48.153846 \quad \sum_{i=1}^{13} X_{31}^2 = 33050$$

$$\sum_{i=1}^{13} Y_i = 1240.5 \qquad \bar{Y} = 95.423077$$

$$\sum_{i=1}^{13} X_{2i}X_{3i} = 4922 \qquad \sum_{i=1}^{13} Y_i^2 = 121088.09$$

$$\sum_{i=1}^{13} X_{2i}Y_i = 10032 \qquad \sum_{i=1}^{13} X_{3i}Y_i = 62027.8$$

$$x'x = \begin{bmatrix} 415.230769 & 251.076923 \\ 251.076923 & 2905.692310 \end{bmatrix}$$

$$x'\underline{y} = \begin{pmatrix} 775.961538 \\ 2292.953850 \end{pmatrix}$$

$$\underline{y}'\underline{y} = 2715.7631$$

Step 2: $(x'x)^{-1} = \begin{bmatrix} 0.002541066 & -0.0002195701 \\ -0.0002195701 & 0.0003631248 \end{bmatrix}$

Step 3: $B_2 = 1.468306$

$B_3 = 0.6622505$

$B_1 = 52.57735$

The regression equation is

$$Y = 52.57735 + 1.468306\ X_2 + 0.6622505\ X_3$$

Step 4: $\underline{e}'\underline{e} = 57.9045$

$R^2 = 0.9786784$

$w^2 \doteq v^2 = 5.790450$

Step 5: $\text{var}(B_1) = 5.226595$

$\text{cov}(B_1, B_2) = -0.04856520$

$\text{cov}(B_1, B_3) = -0.09176431$

$\text{var}(B_2) = 0.01471392$

$\text{cov}(B_2, B_3) = -0.001271410$

$\text{var}(B_3) = 0.002102656$

Step 6: Confidence limits for Y when $X_2 = 6$ and $X_3 = 50$ are

$$94.49971 \pm 2.506258\ t$$

The point estimate for Y is 94.49971, and t is from a Student's t distribution with ten degrees of freedom. For example, the 95 percent confidence interval (r = 0.05; t = 2.228 from the 97.5 percent column of a t-distribution table) is

(88.91577,100.0837)

Step 7: Finally, with $Y' = 120$, when $X_3 = 50$,

$$f = 2.155922 - 0.01471392\ t^2$$

$$g = -100.7555 + 0.2242714\ t^2$$

$$h = 1177.185 - 7.097254\ t^2$$

The resulting limits for different significance levels are given in the last two columns of Table 2. Since the point estimate of X_2 is 23.36715,

TABLE 2 - LIMITS FOR X_2 IN NUMERICAL EXAMPLE

r	t	Z_e	Z_u
0.001	4.587	15.06735	36.94749
0.01	3.169	17.39803	31.65355
0.05	2.228	19.03294	28.80569
0.10	1.812	19.78620	27.66996
0.20	1.372	20.60728	26.53686
0.50	0.700	21.91731	24.92266

$Y' = 120$ lies within the indicated confidence interval for Y so long as X_2 is between Z_e and Z_u. For example, $Y' = 120$ lies within a 95 percent confidence interval for Y so long as X_2 is between 19.03294 and 28.80569.

Similarly, with $Y' = 120$, when $X_2 = 6$,

$$f = 0.4385757 - 0.002102684\ t^2$$

$$g = -77.63273 + 0.1987855\ t^2$$

$$h = 3435.462 - 10.96396\ t^2$$

The resulting limits at different significance levels are given in Table 3. The point estimate of X_3 is 88.50550. Hence, Y' = 120 lies

TABLE 3 - LIMITS FOR X_3 IN NUMERICAL EXAMPLE

r	t	Z_e	Z_u
0.001	4.587	69.75458	116.50926
0.01	3.169	75.06417	106.11853
0.05	2.228	78.78176	100.23987
0.10	1.812	80.48814	97.84187
0.20	1.372	82.34159	95.42050
0.50	0.700	85.28138	91.92384

within the confidence interval for Y so long as X_3 is between Z_e and Z_u as indicated in Table 3.

7. SUMMARY

Standard references on multiple linear regression give explicit formulas for determination of the regression parameters, the estimated covariance structure of these parameter estimates, and interval estimates for the dependent variable (about a value predicted from the regression equation), given values for the independent variables. When the regression is performed in deviation from, the constant term must be calculated from an equation which is generally given in such references. However, formulas for calculation of the portions of the covariance matrix associated with the constant term are generally missing. Such formulas are derived in this report.

In addition, formulas are derived for limits, similar to confidence limits, on the value of an independent variable, given values for the dependent variable and each of the other independent variables. The proper interpretation of such limits is also given.

The formulas developed here have been utilized in a pair of computer programs, one a batch program and the other an interactive program. Both

programs use the IMSL routines to perform a stepwise multiple linear regression. A future report will discuss the IMSL routines used and the programs themselves.

REFERENCES

1. Acton, F.S., "Analysis of Straight-Line Data," John Wiley and Sons, Inc., New York (1959).

2. Draper, W.S. and H. Smith, "Applied Regression Analysis," John Wiley and Sons, Inc., New York (1966).

3. Johnston, J., "Econometric Methods," Second Edition, McGraw-Hill Book Co., New York (1972).

4. "IMSL Library 3 Reference Manual," Fifth Edition, International Mathematical and Statistical Libraries, Houston, Texas (Nov 1975).

5. Dorn, W.S. and D.D. McCracken, "Numerical Methods with FORTRAN IV Case Studies," John Wiley and Sons, Inc., New York (1972).

INITIAL DISTRIBUTION

Copies	
3	NAVPGSCOL 1 Library 1 J. Hartman, Code 55Hh 1 D. Barr, Code 55Bn
1	NAVSEA Code 03F (B. Orleans)
5	NAVSEC 6107E1
12	DDC
1	MSC P. Morfogenis, Code M-62a5

CENTER DISTRIBUTION

Copies	Code	Name
1	18	G. Gleissner
1	1805	E. Cuthill
2	1809.3	D. Harris
1	184.1	H. Feingold
1	187	J. Spurway
1	187	M. Zubkoff
20	187	G. Humfeld
10	5214.1	Reports Distribution
1	522.1	Unclassified Lib (C)
1	522.2	Unclassified Lib (A)

DTNSRDC ISSUES THREE TYPES OF REPORTS

1. DTNSRDC REPORTS, A FORMAL SERIES, CONTAIN INFORMATION OF PERMANENT TECHNICAL VALUE. THEY CARRY A CONSECUTIVE NUMERICAL IDENTIFICATION REGARDLESS OF THEIR CLASSIFICATION OR THE ORIGINATING DEPARTMENT.

2. DEPARTMENTAL REPORTS, A SEMIFORMAL SERIES, CONTAIN INFORMATION OF A PRELIMINARY, TEMPORARY, OR PROPRIETARY NATURE OR OF LIMITED INTEREST OR SIGNIFICANCE. THEY CARRY A DEPARTMENTAL ALPHANUMERICAL IDENTIFICATION.

3. TECHNICAL MEMORANDA, AN INFORMAL SERIES, CONTAIN TECHNICAL DOCUMENTATION OF LIMITED USE AND INTEREST. THEY ARE PRIMARILY WORKING PAPERS INTENDED FOR INTERNAL USE. THEY CARRY AN IDENTIFYING NUMBER WHICH INDICATES THEIR TYPE AND THE NUMERICAL CODE OF THE ORIGINATING DEPARTMENT. ANY DISTRIBUTION OUTSIDE DTNSRDC MUST BE APPROVED BY THE HEAD OF THE ORIGINATING DEPARTMENT ON A CASE-BY-CASE BASIS.